学术图书选题知识图谱研究

摘　要：知识图谱是一种结构化的语义知识库，以图形化的方式描述知识资源。通过可视化的知识图谱可以清晰地展示学术图书选题相关信息间的关联关系以及整体知识脉络，为优化选题相关内容、有效分析挖掘出好的选题提供支撑。本文对学术图书选题知识图谱进行了研究，提出了学术图书选题知识图谱的表示和构建方法，为学术图书选题策划提供辅助决策支撑。

关键词：学术图书选题；知识图谱；辅助决策

中图分类号：G350　　　文献标识码：A

文章编号：1671-0134（2019）12-078-03　　　DOI：10.19483/j.cnki.11-4653/n.2019.12.022

本文著录格式：吴娜达，李彩珊，哈爽 . 学术图书选题知识图谱研究 [J]. 中国传媒科技，2019（12）：78-80.

文 / 吴娜达　李彩珊 *　哈爽

引言

在出版领域，选题是指经过多方面分析、考量而选中主题后拟实施的出版项目。[1] 传统的选题策划多凭借编辑的经验，数据的处理多采用孤立的方式，一般对每一项内容进行单独分析，数据存储方式简单，保存时期短，无法实现信息的精细化管理和多元化、多角度的延伸，没有充分挖掘采集数据的价值。在当下移动互联网快速发展和迅速普及的时代，必须通过一定的手段对数据进行重构和应用，才能在数据内容呈爆炸式增长的态势下快速地获得有价值的信息。

本文对学术图书选题知识图谱的表示、构建等进行研究，以期通过构建的学术图书选题知识图谱对编辑策划学术图书选题提供有效的辅助决策。

1. 学术图书选题现状分析

一般来说，选题来源于编辑在信息采集过程中产生的某种意向或愿望，通过周密分析、研究主客观条件、多方论证等逐步形成方案。选题信息的采集是选题策划中重要的步骤，是选题发现、策划、验证、论证的基础。

1.1 学术图书选题策划采集的信息类型

学术图书选题策划采集的信息一般包含：社会信息、学科信息、出版信息、市场（读者）信息、读者信息 5 部分。[2]

（1）社会信息包含学科所涉及的中央和地方政府的法规、政策、白皮书、公开信息等。

（2）学科信息包含学科的范围和主要内容、国内外发展情况、前沿发展方向和重点方向、重点科研项目、研究课题、学科领军人物、学术成果及褒奖。

（3）出版信息包含两个方面：一方面指本出版社同类选题图书的品种、数量、作者、销售状况等信息；另一方面指同行，即其他出版社出版该类选题图书的品种、数量、作者、销售状况等信息。

（4）作者信息包含作者的学科背景、职务职称、研究方向、主要工作、已申请课题、著作情况等。

（5）读者具有个性化特征，主要信息包含读者基本信息（年龄、职业等）、购买力、读者实际需求、潜在需求、图书市场分布、图书市场反馈等。

1.2 采集信息存在的问题

目前，学术图书选题策划过程中获取的信息多以文档、表格及少量数据库的形式存储，存在一些问题，主要如下。

1.2.1 信息异构

在图书选题信息采集过程中，获取的信息的来源广泛，数据结构不全相同，这给数据的融合、存储带来了巨大的困难。

1.2.2 信息冗余

不同来源的信息组合难度高、优势互补性差，信息的完整性不高。

信息存在大量的冗余与噪声，信息的准确度有待考究。

1.2.3 信息孤立

空间上不连续：关注的信息内容不能在数据上体现事件与事件之间的联系。

时间上不连续：关注的信息内容时间持续性短。

1.2.4 信息静止

不能有效利用已有信息进行发现与预测。

1.2.5 信息可视化困难

数据不能以多种形态表现，使其更直观、更易于理解。

以上问题导致图书选题信息存储难、检索难、重复利用与共享难。

在传统的图书选题的信息采集过程中，多是通过编辑的搜集，进行简单存储（多以文档、表格及少量数据库的形式存储），依赖人力主观对读者、作者、内容、营销等方面的信息进行思考和分析，形成选题策划方案。传统的学术图书选题方式主观性因素大，信息分析不够精确，可靠性和科学性不高，且信息检索、重复利用及共享困难。因此，本文对学术图书选题知识图谱知识表示和构建方法进行研究，通过对信息数据进行重构，使对信息的认识更加清晰、立体，并以期通过推理等算法实现模拟大脑综合分析信息的能力，辅助图书选题进行有效决策。

2. 学术图书选题知识图谱的定义

学术图书选题知识图谱旨在构建一张巨大的语义网络图，用以描述学术图书选题策划过程中存在的各类实体及其关系。图的节点表示实体，图的边表示关系。也可以认为学术图书选题知识图谱是一个大规模的知识库，为学术图书选题中涉及的复杂数据提供有效的存储、检

* 本文通讯作者

索及可视化，为学术图书选题策划提供可靠、清晰的信息及脉络。

目前，学术图书选题知识图谱的数据来源主要基于第 2 节所述的社会信息、学科信息、出版信息、市场（读者）信息、读者信息 5 个方面相关信息，并可根据实际需求进行扩展。5 个方面信息涉及的数据类型主要有 3 类。

2.1 结构化数据

主要指关系数据库中表、excel 表以及其他具有结构的数据。在学术图书选题知识图谱的构建中，其主要来源于出版社各级系统数据库及合作商可提供的数据库等。

2.2 无结构化数据

在学术图书选题知识图谱构建中，主要指纯文本资料，例如硕博论文、报纸、会议的图像和声音等数据。

2.3 半结构化数据

主要指介于结构化数据和无结构化数据之间，通常的 XML、HTML 等相关网页均属于半结构化数据。半结构化数据在学术图书选题构建中，主要来源于各类网站获取的信息，例如从电商网站（图书商城）的 XML 中获取的图书信息（图书名称、编辑推荐、作者简介等）和图书市场信息（评价星级、评价时间、评价人地理位置等）、从工业和信息化部网站的 XML 中获取的公开信息等。

学术图书选题知识图谱的表示与构建参考一般知识图谱的构建过程，通过三元组对学术图书选题知识图谱进行表示，通过图形数据库 Neo4j 的规范设计存储模式及构建学术选题知识图谱。

3. 学术图书选题知识图谱的表示

学术图书选题知识图谱的结构由节点集合和边集合构成，形式化表示如式（1）：

ATS_KG={<ATS_N>，<ATS_R>}（1）

其中，<ATS_N> 表示学术图书选题的节点集合，节点是学术图书选题信息中的各种实体，例如作者、书籍；<ATS_R> 表示学术图书选题的边集合，可表达为如式（2）：

<ATS_R >={<ATS_T>，<ATS_D>，<ATS_G>}（2）

其中，<ATS_T> 表示关系的类型集合，例如“作者—书籍关系”“书籍—出版社关系”；<ATS_D> 表示关系的方向集合，例如“作者—> 书籍”“书籍 <—出版社”；<ATS_G> 表示三元组集合，通过三元组表达语义关系，每一个三元组表示一个事实，可表示为如式（3）：

<ATS_G>={（ATS_N_1，ATS_T_1，ATS_N_2）}（3）

其中，式（3）的含义是，ATS_N_1 与 ATS_N_2 分别表示不同的节点（实体），ATS_T_1 表示 ATS_N_1 与 ATS_N_2 之间的语义关系，方向是由 ATS_N_1 指向 ATS_N_2。例如存在事实：作者李杰，著作《工业大数据》，可用三元组（李杰，作者—书籍关系，《工业大数据》）进行表示。

4. 学术图书选题知识图谱的构建

学术图书选题知识图谱的构建主要有 2 个步骤，包括学术图书选题数据库存储模式设计、利用图形数据库构建知识图谱。如图 1 所示。

4.1 数据库存储模式设计

梳理学术图书选题相关信息，对实体及其之间的关系进行规范的建模，并给出明确的定义。结合第 3 节中学术图书选题信息涉及的 3 种数据类型和学术图书选题知识图谱的表示，对学术图书选题知识图谱数据库存储模式进行设计。对 3 种数据类型分别进行介绍。

（1）结构化的数据主要来自关系型数据库（例如 MySql、SQL Server）、Excel 等，主要方法是通过分析表的信息和字段信息，抽取出关系模式，设计转化规则，建立学术图书选题知识图谱图数据库的表结构。

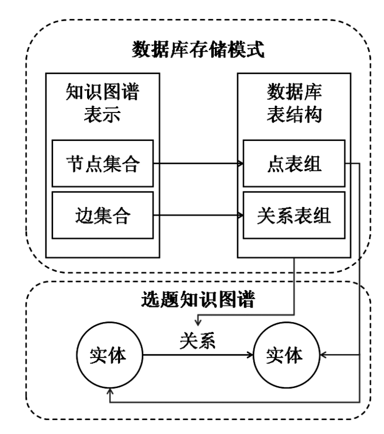


图 1　学术图书选题知识图谱的构建

关系型数据库 MySql 中存在表名为“作者信息”的表，见表 1。

表 1　作者信息

ID	姓名	年龄	单位	职务	研究方向	出版书籍	
1	刘某	34	高校 1	教授	无人驾驶、强化学习	《A》	…
2	张某	56	研究所 1	研究员	机器视觉	《B》	…
3	郭某	48	高校 2	教授	无线通信	《C》	…
…	…	…	…	…	…	…	…

结合学术图书选题知识图谱的表示可抽象得到点集合和关系集合。

点集合

姓名 ={ 刘某，张某，郭某，…}；

年龄 ={34，56，48，…}；

单位 ={ 高校 1，研究所 1，高校 2，…}；

职务 ={ 教授，研究院，…}；

研究方向 ={ 无人驾驶，强化学习，机器视觉，图像处理，无线通信，…}；

出版书籍 ={《A》，《B》，《C》，…} 等。

边集合

作者－单位关系 ={< 作者－单位关系，作者—> 单位，（刘某，作者－单位关系，高校 1）>，< 作者－单位关系，作者—> 单位，（张某，作者－单位关系，研究所 1）>，< 作者－单位关系，作者—> 单位，（郭某，作者－单位关系，高校 2）>…}；

作者－研究方向关系 ={< 作者－研究方向关系，作者—> 研究方向，（刘某，作者－研究方向关，无人驾驶）>，< 作者－研究方向关系，作者—> 研究方向，（刘某，作者－研究方向关，强化学习）>，< 作者－研究方向关系，作者—> 研究方向，（张某，作者－研究方向关，机器视觉）>，< 作者－研究方向关系，作者—> 研究方向，（郭某，作者－研究方向关，无线通信）>，…} 等。

由于篇幅限制，此处不一一列举存在的点集合和边集合。

通过节点集合和关系集合进行学术图书选题知识图谱存储模式的设计。节点集合映射为学术图书选题知识

图谱图数据库存储模式中的点表组，边集合映射为图数据库存储模式中的关系表组。学术图书选题知识图谱的图数据库存储模式表示为实际存储的点表组和关系表组。点表组可以很好地保留层级结构，同时，可以有效扩展实体数据。关系表组由三元组表示，存放事实，关系表组可以很好地保留语义关系的层级关系，同时，可以对实体之间的其他语义关系进行有效存储与拓展。

以作者 – 研究方向关系涉及的节点集合和边集合为例，映射为相应的点表组和关系表组，作者点表见表 2，研究方向点表见表 3，作者 – 研究关系表见 4。

表 2　作者点表

ID	节点 1	标签
Name_1	刘某	姓名
Name_2	张某	姓名
Name_3	郭某	姓名
…	…	…

表 3　研究方向点表

ID	节点 2	标签
Research_1	无人驾驶	研究方向
Research_2	强化学习	研究方向
Research_3	机器视觉	研究方向
Research_4	无线通信	研究方向
…	…	…

表 4　作者 – 研究关系表

节点 1_ID	关系 1	节点 2_ID
Name_1	作者 – 研究关系	Research_1
Name_1	作者 – 研究关系	Research_2
Name_2	作者 – 研究关系	Research_3
Name_3	作者 – 研究关系	Research_4
…	…	…

（2）无结构化数据根据语义关系的层级和结构进行自定义。

获取一段文本信息“2016 的政策热点关键词有大数据、物联网、云计算、人工智能、5G 技术等，大数据的学科方向可分为大数据采集与预处理、大数据存储与管理、大数据计算模式、大数据分析与挖掘、大数据可视化、大数据安全、大数据系统与内存、大数据处理平台、大数据网络等，其中，工业大数据方向的图书市场上有《工业大数据》，作者是李杰，出版单位是机械工业出版社。”

结合学术图书选题知识图谱的表示可抽象得到点集合和关系集合。

点集合

政策热点点集合 ={ 大数据，物联网，云计算，人工智能，5G 技术，…}；

大数据学科方向点集合 ={ 大数据采集与预处理，大数据存储与管理，大数据计算模式，大数据分析与挖掘，大数据可视化，大数据安全，大数据系统与内存，大数据处理平台，大数据网络，数据算法基础，工业大数据，…} 等。

边集合

热点 – 学科边集合 ={< 热点学科关系，热点—> 学科，(大数据，热点学科关系，工业大数据)>，< 热点学科关系，热点—> 学科，（大数据，热点学科关系，大数据存储与管理 ）>，…}。

各类关系通过逻辑及实际应用进行解析定义，此处不一一列举。

通过节点集合和关系集合进行学术图书选题知识图谱图数据库存储模式的设计方法与结构化数据，此处不再列举。

（3）半结构化数据介于结构化数据和非结构数据之间，一般在 XML 文件中得到相关的数据层级关系，根据实际情况，结合结构化数据和非结构化数据的处理方法即可设计得到学术图书选题知识图谱图数据库模式。

4. 2 利用图数据库构建知识图谱

基于设计的图数据库存储模式，利用图形数据库进行学术图书选题知识图谱的构建。从最近几年的发展情况可以看到基于图的存储方式在整个数据库存储领域的飞速发展，其中，Neo4j 在整个图存储领域中占据 NO.1 的地位。Neo4j 是一个高性能的图数据库，可以在 Java 和 Python 等平台上使用。本文通过 Java API 将图数据存储模式中的点表组与关系表组中的学术图书选题数据写入 Neo4j 图数据库中，构建学术图书选题知识图谱。利用 Neo4j 图数据库构建知识图谱可参考 The Neo4j Graph Platform，主要方法是将设计的点表组和关系表组，利用 Cypher 语句写入 Neo4j 图形数据库，得到知识图谱。

以人民邮电出版社国之重器系列图书为例给出部分知识图谱展示示意图，如图 2 所示。图 2 中对人民邮电出版社国之重器系列图书的相关信息进行了部分示意。

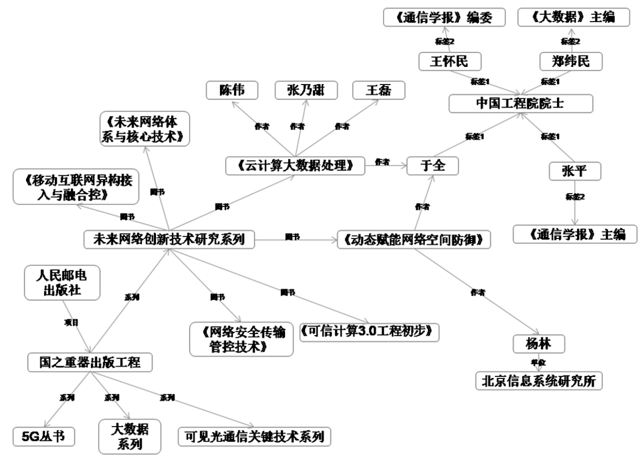


图 2　学术图书选题部分知识图谱示意

结语

本文通过对学术图书选题策划中存在的问题进行分析，对学术图书选题知识图谱的表示和构建方法进行研究，以期通过构建知识图谱解决目前学术图书选题策划中存在的问题，并为学术图书选题策划提供有效的辅助决策支撑。目前，因数据量的限制，本文的知识图谱规模较小，后续将继续丰富学术图书选题知识图谱，并以期通过知识推理等方法推荐辅助学术图书选题策划。媒

参考文献

[1] 全国出版专业职业资格考试办公室 . 出版专业基础知识 [M]. 上海：上海辞书出版社，2004.

[2] 张伯熙 . 图书选题策划前期信息采集 [J]. 中国传媒科技，2013（4）：226-227.

（作者单位：北京信通传媒有限责任公司）